

Threats to Validity in Software Engineering – hypocritical paper section or essential analysis?

Patricia Lago*
p.lago@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Qunying Song*
qunying.song@cs.lth.se
Lund University
Lund, Sweden

Per Runeson*
per.runeson@cs.lth.se
Lund University
Lund, Sweden

Roberto Verdecchia*
roberto.verdecchia@unifi.it
University of Florence
Florence, Italy

International Symposium on Empirical Software Engineering and Measurement (ESEM), 2024, pp. 314 - 324

背景 | 妥当性への脅威(Threats To Validity)

ソフトウェア工学研究では、実証的評価が一般的に^[28]

- 厳密で透明性のある評価が必要

妥当性への脅威 (TTV)

- 論文に掲載される、研究結果の信頼性を示す重要な要素
- 一般的に、論文のTTVセクションでの議論が推奨される^[7]

TTVセクションの主な目的^[7]

- 研究結果の位置づけの明確化
- 脅威の緩和戦略の報告

[28] Li Zhang, Jia-Hao Tian, Jing Jiang, Yi-Jun Liu, Meng-Yuan Pu, and Tao Yue. 2018. Empirical Research in Software Engineering — A Literature Survey. Journal of Computer Science and Technology 33, 5 (2018), 876–899.

[7] Andreas Jedlitschka and Dietmar Pfahl. 2005. Reporting guidelines for controlled experiments in software engineering. In International Symposium on Empirical Software Engineering (Noosa Heads, QLD, Australia). IEEE, 10 pp.–.

問題 | TTVセクションの存在と品質

TTV分析が形骸化している

- 形式的に後付けされる形となっており、「偽善的なセクション」であると批判される
- TTVセクションの弱点や懸念について、長らく議論されている^[5]

一方、TTVの現状に関する体系的な調査はほとんど行われていない

- 関連研究
 - 脅威間の優先順位やトレードオフに関する議論^[15]
 - 構成妥当性に関する分析^[24]

問題

TTVの現状を示す定量的な証拠の不足
改善への具体的な指針の欠如

[5] Robert Feldt and Ana Magazinius. 2010. Validity Threats in Empirical Software Engineering Research – An Initial Survey. In Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE'2010). Knowledge Systems Institute Graduate School, 374–379.

[15] Martin P. Robillard, Deeksha M. Arya, Neil A. Ernst, Jin L. C. Guo, Maxime Lamothe, Mathieu Nassif, Nicole Novielli, Alexander Serebrenik, Igor Steinmacher, and Klaas-Jan Stol. 2024. Communicating Study Design Trade-offs in Software Engineering. ACM Transactions on Software Engineering and Methodology 33, 5 (2024), 1–10.

目的

ソフトウェア工学研究におけるTTV実践の現状を体系的に把握・分析する

- より透明で有用なTTV分析・報告に向けた改善提案へ

調査目標

- 既存の論文のTTVの特徴の把握
- 時間的な改善傾向の特定
- コミュニティのTTV実践状況の評価

アプローチ

調査対象

- 2014~2023のICSE論文
- Distinguished Paper Awardを受賞した計91論文

8つの研究設問それぞれについて論文を分類

設問

Q0: 研究タイプの分布

Q1: TTV議論の評価

Q2: ガイドライン使用状況の調査

Q3: 使用される脅威カテゴリーの特定

Q4: TTVの分類適合性の評価

Q5: 事前分析の有無確認

Q6: 限界議論の調査

Q7: トレードオフ議論の分析

4人の著者による分担評価

調査手順

①分類・評価基準の確立

- 最新の年の3論文を参考に、分類基準の統一などを行う
 - 例：Q3 “使用される脅威カテゴリーの特定” について、
 - カテゴリ：内部妥当性, 外部妥当性, 構成概念妥当性, 結論の妥当性 などの有無を判断

②データ収集・分類

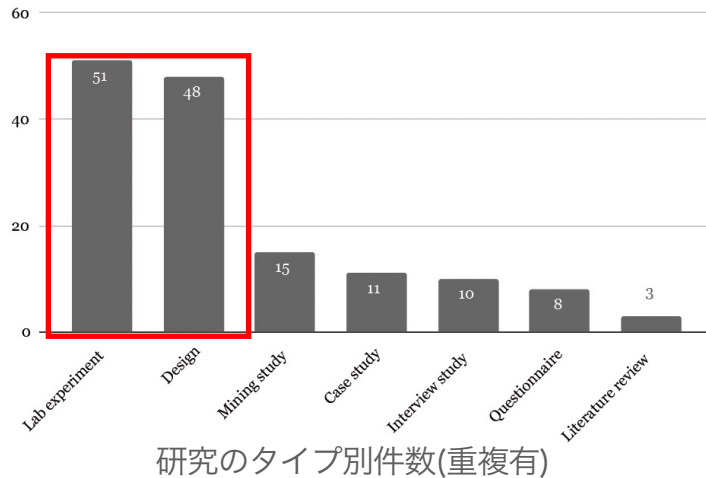
- 91論文を4人で分担
- 論文の時系列バイアスを避けた割り当て
- 評価品質保証のため、3バッチに分けた段階的な評価

③分析・検証

- 分類結果の分析
- 境界事例の合議による判断

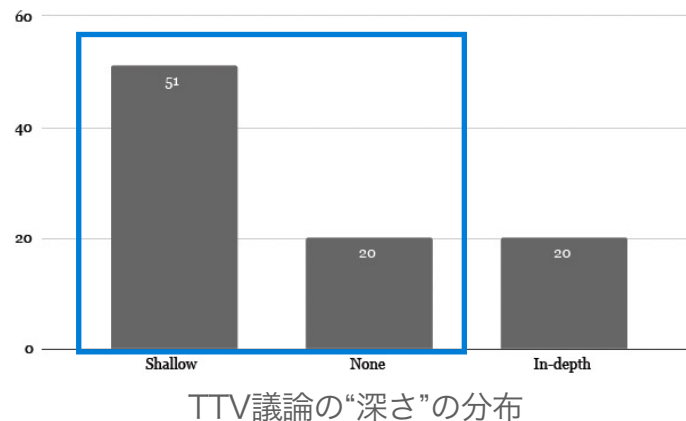
結果① | TTVの反映状況

Q0.研究タイプ(手法)の分布



- ・ 実験室実験：51件
- ・ デザイン研究：48件
- ・ 複数手法の組み合わせ：55件

Q1.TTV議論の深さの評価

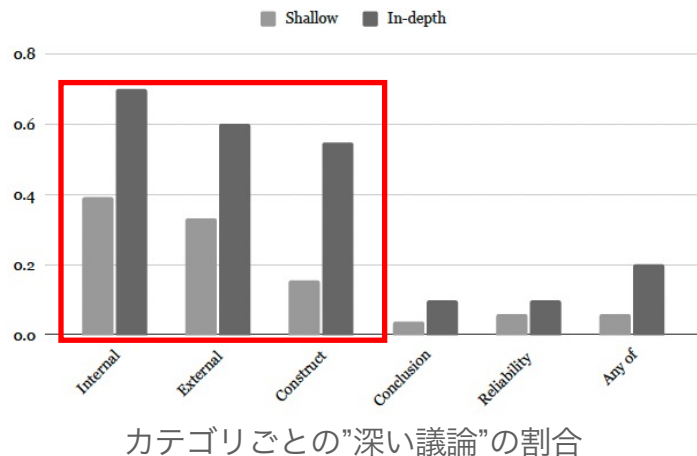
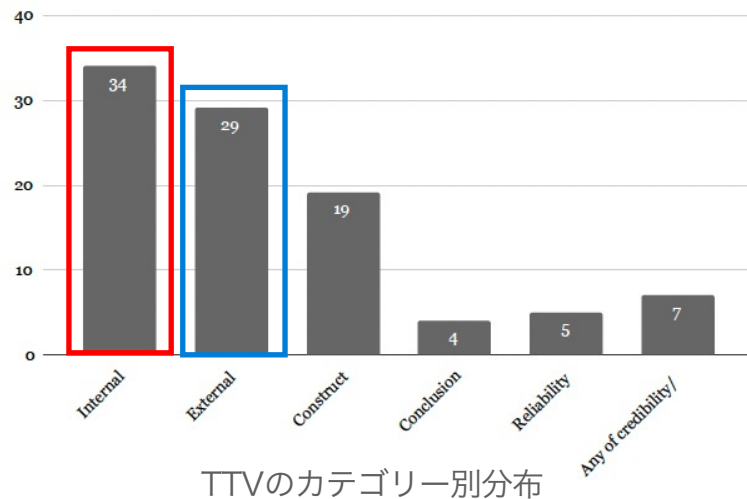


- ・ 深い議論：20件
- ・ 浅い議論：51件
- ・ 議論無し：20件

⇒ トップ論文でも78%が不十分

結果② | TTVの内容分析

Q3.使用されるTTVカテゴリー



- 内部妥当性：34件
- 外部妥当性：29件
- 構成概念妥当性：19件
- 大多数が複数のカテゴリーを組み合わせで議論

“深い議論”を行う論文の特徴として、主要3カテゴリーを包括的・積極的に議論

結果③

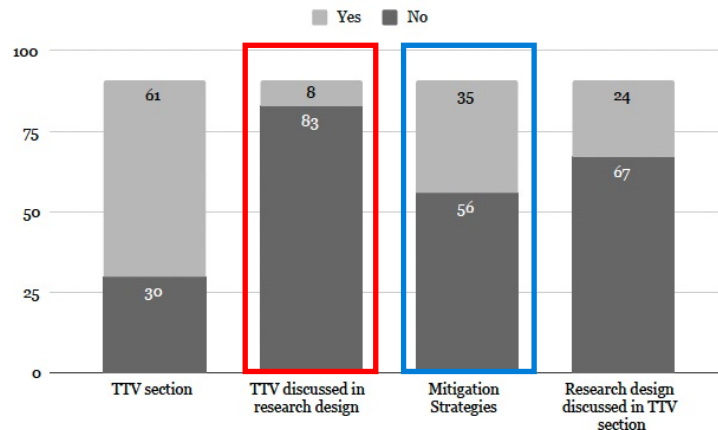
Q4. 掲載したTTVのカテゴリは研究に適合しているか

	適合	不適合
2014-2018	39.5%(15本)	60.5%
2019-2023	26.4%(14本)	73.6%
全体	31.9%(29本)	68.1%(62本)

- ・ 大多数のTTVが研究に不適合
- ・ 10年間で改善傾向が見られない

研究内容に対して、適切なTTVカテゴリを選択できている割合

Q5. 事前のTTV分析の兆候はあるか



事前のTTV分析の兆候に関する分布

- ・ 研究設計段階でTTVはほとんど考慮されていない
- ・ 大多数の論文でTTV緩和戦略が議論されていない

⇒ TTVが形式的な後付けと扱われる

考察 | 不十分なTTVを受けて

論文の標準的構造を重視し過ぎている

- 本来の目的（複製・拡張・転移可能性）への注目が不足

TTVを深く議論しない理由

- 査読での懸念
 - 正直な限界議論が不利に働くリスク，建設的でない批判への懸念
- 論文上の制約
 - 論文のスペース制約
 - TTVと研究貢献の記述量のトレードオフ

解決への示唆

- 問題はコミュニティ全体で広く発生
- 個人ではなく，コミュニティ全体での理解向上が必要

TTV考慮事項の提案（一部）

設計段階

- 研究目標に応じた妥当性カテゴリーの選択
- TTV同士のトレードオフを意識した設計
- 研究手法選択時のTTV考慮

実験段階

- 設計選択がTTVに与える影響の考慮
- 分野内で確立された実践手法の活用

分析・報告段階

- TVの詳細な分析の実施
- 査読者ではなくユーザー視点での記述
- バランスの取れたTTV/限界の提示

総括・所感

研究の主要な発見

- TTVの形式的な扱いが依然として主流である(78%が不十分)
- 10年間で改善傾向が見られない
- 体系的なアプローチが欠如

展望

- TTVの体系的な評価方法の確立
- コミュニティ全体での継続的な改善

所感

- 分析の判断基準をもう少し詳しく知りたかった
- Q1などにおいて、実証研究ではない論文など、TTVセクションの必要性が低い論文の考慮は？

補足1 | 全てのQとその分類

QID	Question	Problem	MID	Metric
Q0	Which type of study is reported?		M0.1	Lab experiment/ Case study/ Mining study/ Interview study / Questionnaire study/ Literature review/ Simulation/ Design
Q1	Is TTV reflected upon generally in the paper?	P1, P5	M1.1	Dominantly "shallow" vs. "in-depth" reflection
Q2	Is a checklist or guideline used?	P1, P6	M2.1	Explicit reference to checklist/guideline?
			M2.2	Which checklist(s)/guideline(s) are referred to?
Q3	Which TTV categories are used?	P2	M3.1	Internal/external/construct/conclusion/reliability – credibility/transferability/dependability/confirmability/other
Q4	Is the TTV categorization fit for the study	P2	M4.1	Is the categorization used matching the current type of study?
Q5	Are there indications of proactive TTV analysis?	P4	M5.0	Is there a TTV (sub-)section?
			M5.1	Is TTV discussed in the research design/methodology section?
			M5.2	Are actions to mitigate TTVs discussed?
			M5.3	Is research design discussed in the TTV section?
Q6	Are limitations discussed?	P7	M6.1	Is there a (sub-)section entitled Limitations?
			M6.2	Are limitations discussed in line with established definitions?
Q7	Are trade-offs between TTV discussed?	P8	M7.1	Are trade-offs discussed in the research design/methodology section?
			M7.2	Are trade-offs discussed in the TTV section?

Table 1: Questions (Qn) and metrics (Mn.m) for the data collection. Problems (Px) refer to Verdecchia et al [24].

補足2 | 全ての提案

CID	Phase	Question / Suggestion	Stakeholders	QIDs
C1	Design	Which categories of validity is most important for the given research goal?	Researchers	Q1
C2	Design	How are TTVs traded against each other?	Researchers	Q7
C3	Design	Which research method(s) are most suitable, given that goal?	Researchers	Q0
C4	Design	Which categories of TTV (or limitations) are most suitable for the chosen type of study?	Researchers	Q3, Q4
C5	Design	Is a suitable checklist or other guidance available for the type of TTV?	Researchers	Q2
C6	Implementation	Which choices are made in the implementation of the research design, that may impact TTV?	Researchers	Q7
C7	Implementation	Are there any established, evidence-based practices within the sub-field that can be used?	Researchers	Q2
C8	Analysis and reporting	Analyze TTV in depth.	Researchers	Q1
C9	Analysis and reporting	Present TTV and limitations in a balanced way. Write for the user of the research, and not for the reviewer.	Researchers	Q5, Q6
C10	Review	Assess TTV of study in relation to its contributions/claims.	Reviewers & Editors	Q0–Q7
C11	Review	Do not punish honesty in TTV/Limitations analysis.	Reviewers & Editors	Q0–Q7
C12	Reading	Assess/cite/use the results in relation to the TTV reporting.	Audiences	Q0–Q7

Table 2: Considerations (indexed by CID) for different stakeholders at different phases of a research study.

補足3 | 妥当性カテゴリ概要

内部妥当性

- この研究結果は、実際に研究で設定した要因によるものと言えるか？

外部妥当性

- この研究結果は他の状況や集団にも当てはまるか？

構成概念妥当性

- 測定しようとしている概念を、実際に反映する形で測定できているか？

結論の妥当性

- このデータに基づいて、結論が正しいと言えるか？

信頼性

- この結果は、何度やっても同じ結果が得られるか？

移転可能性

- この結果は、他の状況や文脈でも当てはまるか？

確認可能性

- 研究結果が、研究者の主観ではなく、データに基づいているか？